Next-generation Sequencing Technology and Data Analysis

- from reads to biology

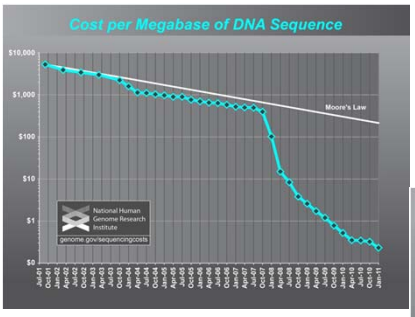
Jason Li Ph.D.

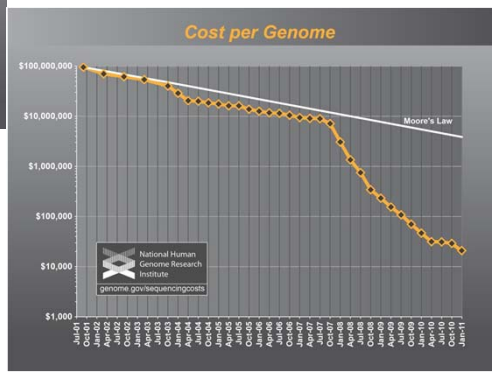
June 8, 2011

Topics

- Technical overview of the Next-generation sequencing (NGS) platforms
- NGS applications
- Experimental design
- Basic data analysis
- Advanced data analysis

Cost of DNA Sequencing

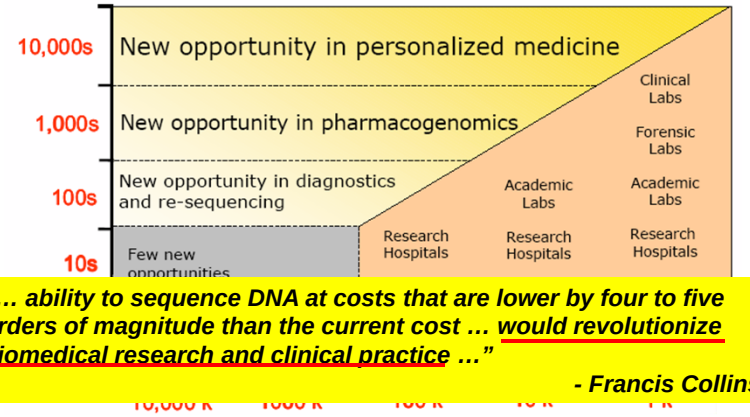




http://www.genome.gov/sequencingcosts

Reduced Cost and Enhanced Speed of DNA Sequencing have Created New Opportunities

Number of Sample in Study



"… ability to sequence DNA at costs that are lower by four to five orders of magnitude than the current cost … would revolutionize biomedical research and clinical practice …"

- Francis Collins

Cost of Re-sequence Human Genome

Technical Overview of the NGS Technologies and Platforms

Major Next Generation Sequencing Platforms

- Roche/454
- Illumina/Solexa
- LifeTech/SOLiD (Lake Nona)
- LifeTech/Ion Torrent (La Jolla)
- Helicos BioSciences
- Pacific Biosciences

Comparison of Next-generation Sequencing Platforms

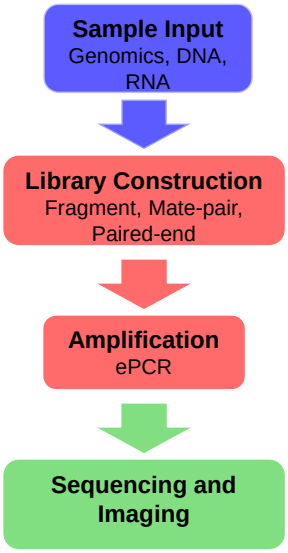
Platform	Library/ template preparation	NGS chemistry	Read length (bases)	Run time (days)	Gb per run	Machine cost (US$)	Pros	Cons	Biological applications	Refs
Roche/454's GS FLX Titanium	Frag, MP/ emPCR	PS	330*	0.35	0.45	500,000	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homo-polymer repeats	Bacterial and insect genome *de novo* assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics	D. Muzny, pers. comm.
Illumina/ Solexa's GA_{II}	Frag, MP/ solid-phase	RTs	75 or 100	4‡, 9§	18‡, 35§	540,000	Currently the most widely used platform in the field	Low multiplexing capability of samples	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Life/APG's SOLiD 3	Frag, MP/ emPCR	Cleavable probe SBL	50	7‡, 14§	30‡, 50§	595,000	Two-base encoding provides inherent error correction	Long run times	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Polonator G.007	MP only/ emPCR	Non-cleavable probe SBL	26	5§	12§	170,000	Least expensive platform; open source to adapt alternative NGS chemistries	Users are required to maintain and quality control reagents; shortest NGS read lengths	Bacterial genome resequencing for variant discovery	J. Edwards, pers. comm.
Helicos BioSciences HeliScope	Frag, MP/ single molecule	RTs	32*	8‡	37‡	999,000	Non-bias representation of templates for genome and seq-based applications	High error rates compared with other reversible terminator chemistries	Seq-based methods	91
Pacific Biosciences (target release: 2010)	Frag only/ single molecule	Real-time	964*	N/A	N/A	N/A	Has the greatest potential for reads exceeding 1 kb	Highest error rates compared with other NGS chemistries	Full-length transcriptome sequencing; complements other resequencing efforts in discovering large structural variants and haplotype blocks	S. Turner, pers. comm.

Metzker M Nat Rev Genet. 2010

SOLiD Sequencing Technology

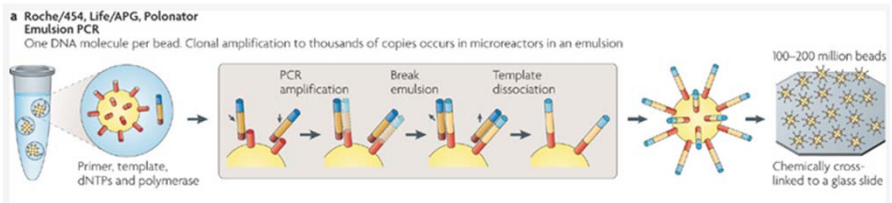
- **SOLiD: Sequencing by Oligonucleotide Ligation Detection**
  - SOLiD system is a highly accurate, massively parallel next-generation sequencing platform.
  - SOLiD v4 system

Next-Generation Sequencing (NGS) Workflow
- Experimental Workflow (Analytical Genomics Core)



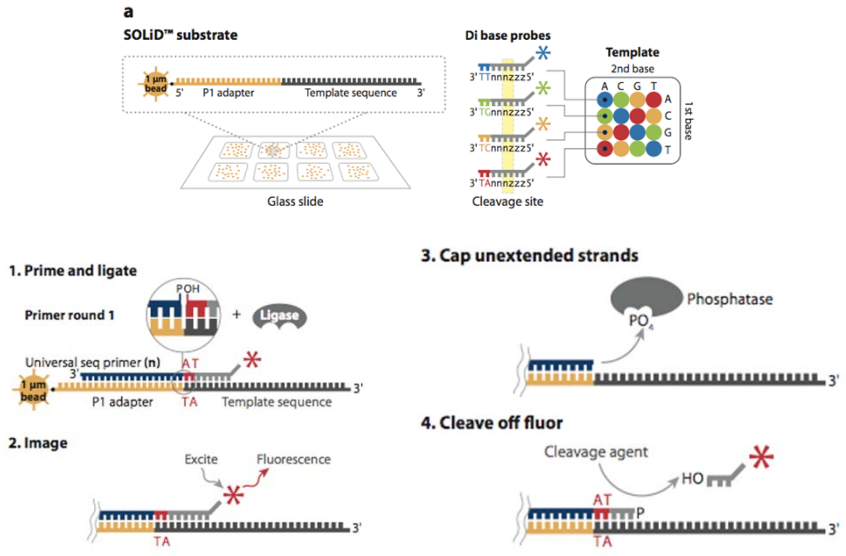
Template Preparation

- **Library construction**
  - *Fragment templates*: randomly shearing genomic DNAs into small sizes of < 1kb
  - *Mate-pair templates*: circularized fragment of >1kb with either single reaction read or two end read (1kb – 10 kb)
  - *Paired-end templates*: linear fragment with ability to sample both ends in separate reactions (200 - 600 bp)
- **Clonally amplified templates**
  - Emulsion PCR: Roche/454; LifeTech/SOLid

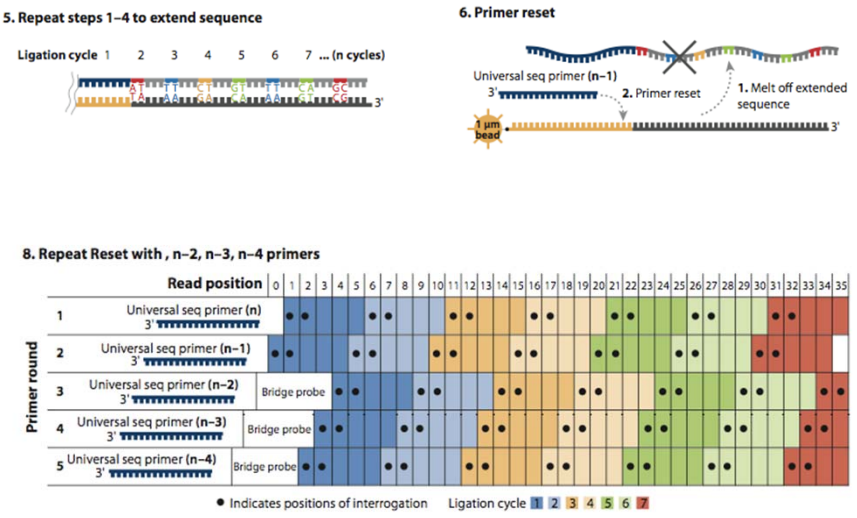


Metzker M Nat Rev Genet. 2010

Sequencing By Ligation Using di-Base
- SOLiD Sequencing Chemistry

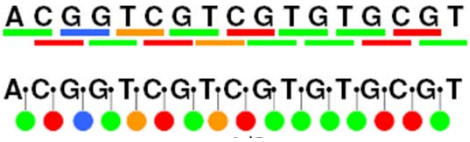


Sequencing By Ligation Using di-Base
- SOLiD Sequencing Chemistry



SOLiD System Color Space di-Base Coding

- Each base is interrogated twice, and the information about each base is included in two adjacent pieces of color space data

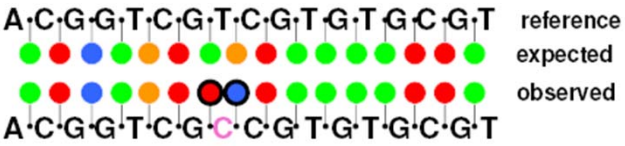


- Unique built-in error checking capability distinguishes between measurement errors and true polymorphisms

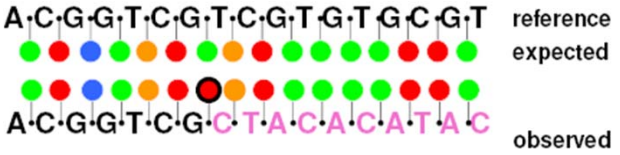
Advantages of di-Base Encoding

- SNP vs Sequencing Error

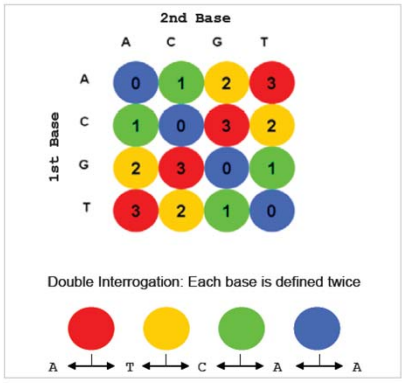
SNP



error



Primary Analysis



Output: *.csfasta

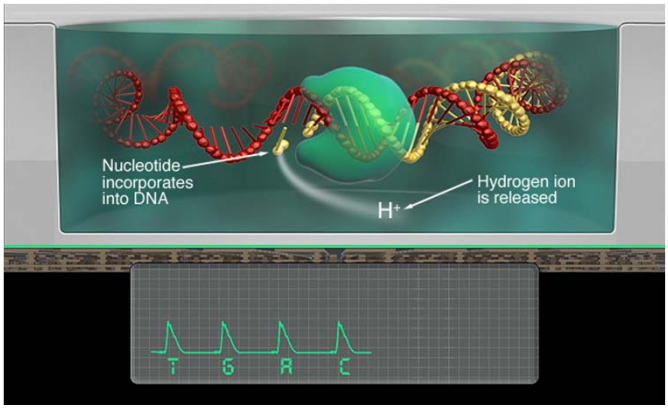
```
>443_1088_005_F3
T32311301011311231133321301012223110
>443_1088_006_F3
T13211113031122103020002220012122101
```

SOLiD v4 Performance Specifications

Library Type	Read Length	Days / Run	Total Tags/Run	Mappable Data
Fragment	1 x 35 bp	3.5 - 4.5	> 700 M	25 - 35 GB
	1 x 50 bp	6 - 8	> 700 M	40 - 50 GB
Paired-End	50 x 25 bp	11 - 13	> 1.4 B	55 - 70 GB
Mate-Paired	2 x 35 bp	8 - 9	> 1.4 B	50 - 70 GB
	2 x 50 bp	12 - 16	> 1.4 B	90 - 100 GB

- **2 slides per instrument run/independent**
- **Ability to barcode and/or divide slides**
- **Very high accuracy data due to di-base encoding**

Ion Torrent Personal Genome Machine (PGM)



- **Semiconductor technology transfers chemical information to digital information.**

Ion Torrent PGM

- **Fastest sequencing workflow**
  - Two-hour sequencing run for up to 200 bp reads
  - Fully prep 8 samples in parallel in less than 6 hours
- **One instrument, your choice of throughput**
  - From 10 Mb to > 1 Gb of highly accurate data
- **Unmatched uniformity of coverage**
  - Simple natural chemistry results in unmatched uniformity of coverage
- **Complete range of applications**
  - Amplicon sequencing, microbial sequencing, RNA-seq, ChIP-seq, methylation, paired-end

NGS Applications

SOLiD™ Applications

Genome

Whole Genome Resequencing
Targeted Resequencing
De Novo Sequencing
Metagenomics

Transcriptome

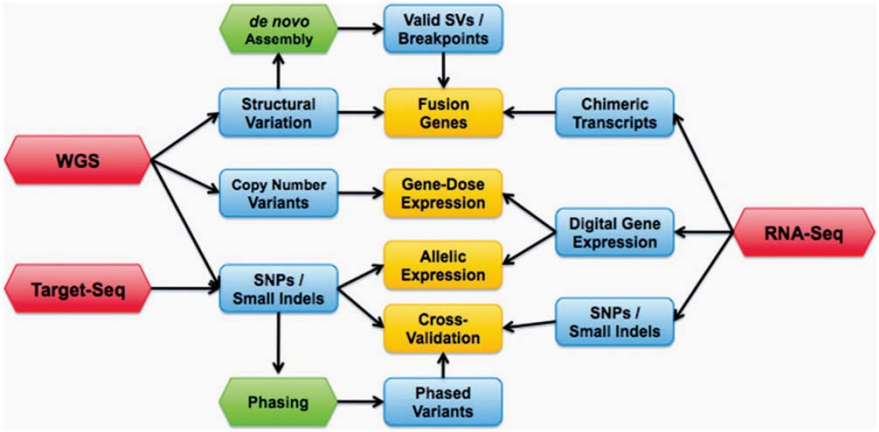
Gene Expression Profiling
Small RNA Analysis
Whole Transcriptome Analysis

Epigenome

Chromatin Immunoprecipitation
Methylation Analysis

applied biosystems
part of life technologies

Intersection of WGS, Target-seq and RNA-seq



Koboldt et al. Brief Bioinform. 2010 11:484

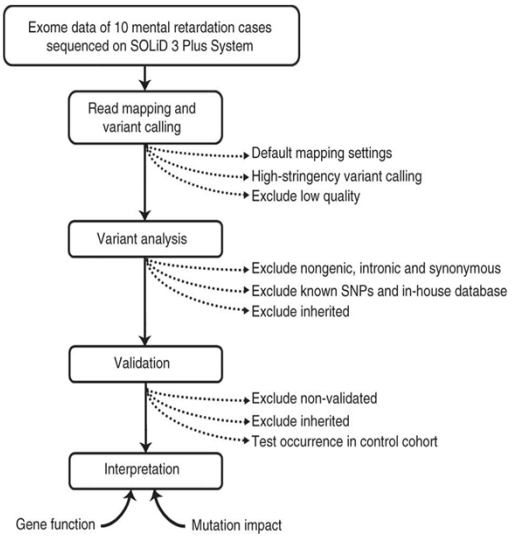
Clinical Applications

- **Detecting Mutations in Disease Influencing Genes**
- **Simultaneously Screening Genetic Diseases**
- **Discovering Disease Influencing Genes**
- **Personalized Sequencing**
- **Improved Cancer Diagnosis and Treatment**
- **Studying Epigenetics**
- **Identifying Structural Variants**
- **Identifying Fusion Genes**
- **Pathogen Detection and Screening**

A *de novo* paradigm for mental retardation

nature genetics

Lisenka E L M Vissers[1,2], Joep de Ligt[1,2], Christian Gilissen[1], Irene Janssen[1], Marloes Steehouwer[1], Petra de Vries[1], Bart van Lier[1], Peer Arts[1], Nienke Wieskamp[1], Marisol del Rosario[1], Bregje W M van Bon[1], Alexander Hoischen[1], Bert B A de Vries[1], Han G Brunner[1,3] & Joris A Veltman[1,3]

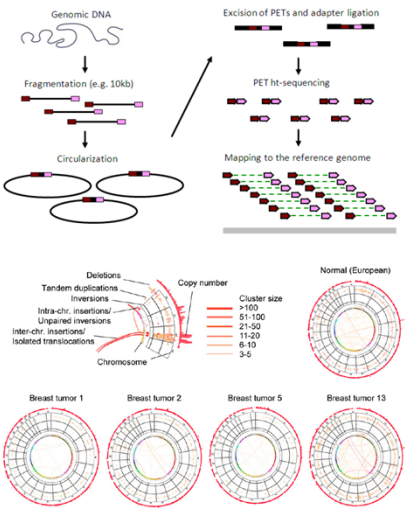


- Family-based exome sequencing approach to test the *de novo* mutation hypothesis in ten individuals with unexplained mental retardation
- Found and validated unique non-synonymous de novo mutations in nine genes. Six of them are likely to be pathogenic.

Comprehensive long-span paired-end-tag mapping reveals characteristic patterns of structural variations in epithelial cancer genomes

Axel M. Hillmer, Fei Yao, Koichiro Inaki, et al.

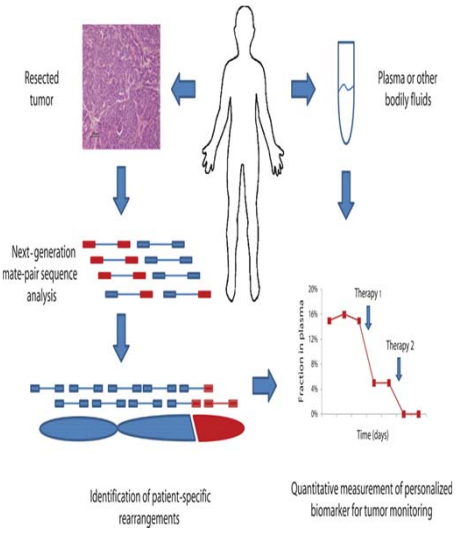
CSH PRESS GENOME RESEARCH



- The authors applied paired-end-ditag (PET) based NGS approach to identify of breakpoints within repetitive or homology-containing regions
- Their approach resulted in a higher physical coverage compared with small insert libraries with the same sequencing effort.
- The PET approach can also be applied to address complex biological questions such as how cancer cells progress and how stem cells maintain their unique properties.

Development of Personalized Tumor Biomarkers Using Massively Parallel Sequencing

Rebecca J. Leary, *et al.*
Sci Transl Med **2**, 20ra14 (2010);
DOI: 10.1126/scitranslmed.3000702

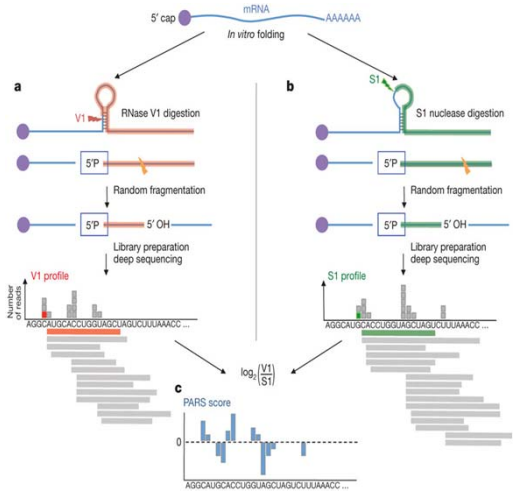


- The authors used SOLiD mate-paired strategy (personalized analysis of rearranged ends (PARE) approach) to identify individualized tumor-specific rearrangements from a small subset of individuals.
- Analyzed 4 colorectal and two breast cancers, and identified an average of nine rearranged sequences per tumor.
- PARE offers an exquisitely sensitive and broadly applicable approach for development of personalized biomarkers to enhance the clinical management of cancer patients

Genome-wide measurement of RNA secondary structure in yeast

nature

Michael Kertesz[1]*†, Yue Wan[2]*, Elad Mazor[1], John L. Rinn[3], Robert C. Nutter[4], Howard Y. Chang[2] & Eran Segal[1,5]

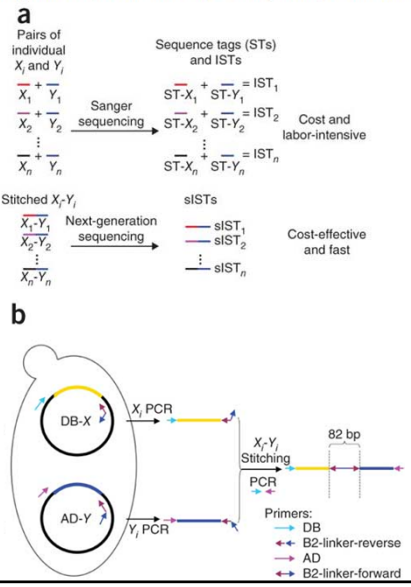


- The authors developed a novel strategy called parallel analysis of RNA structure (PARS) based on deep sequencing fragments of RNAs
- Simultaneous in vitro profiling of the secondary structure of thousands of RNA species at single nucleotide resolution
- Identify over 3,000 distinct transcripts structural profiles in yeast

Next-generation sequencing to generate interactome datasets

478 | VOL.8 NO.6 | JUNE 2011 | NATURE METHODS

Haiyuan Yu, Leah Tardivo, Stanley Tam, Evan Weiner, Fana Gebreab, Changyu Fan, Nenad Svrzikapa,



- NGS has not been applied to protein-protein interactome network mapping
- The new approach called Stitch-seq combines PCR with NGS to generate interactome dataset
- Detect 19% more interactions compared to traditional methods
- Could expand to other binary interaction assays

Applications Completed by Analytical Genomics and Bioinformatics Cores

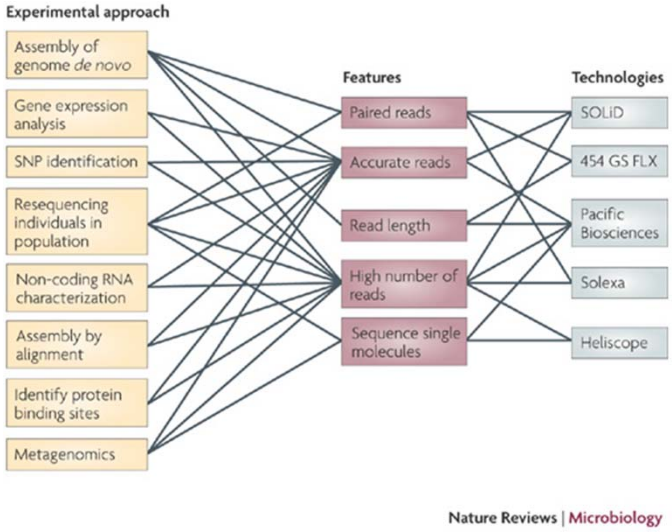
- **Transcriptome**
  - RNA-seq
  - Small RNA-seq
  - Whole transcriptome
- **Genome**
  - De novo sequencing (bacterial)
  - Whole exome capture
  - Whole Genome Re-sequencing
  - Targeted re-sequencing
- **Epigenome**
  - ChIP-seq
  - Methyl-seq (MethylMiner)
- **Innovative Projects**

Experimental Design

Plan Your First NGS Experiment

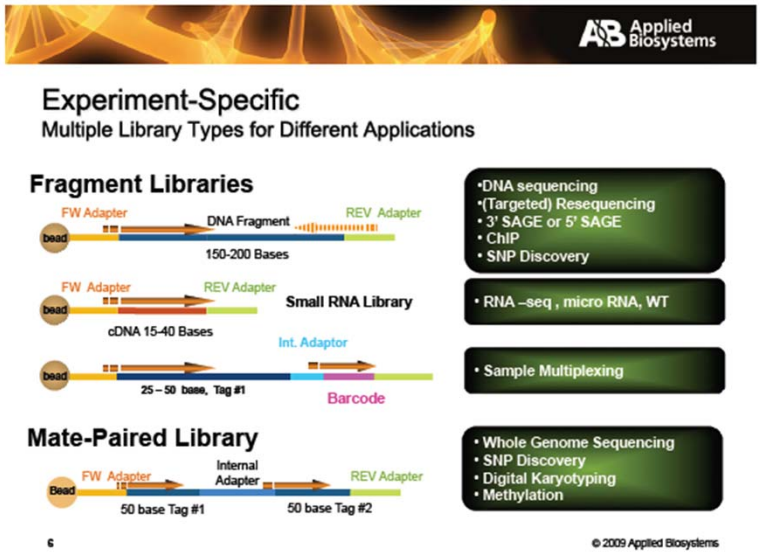
- Sequencing platforms
- Library type
- Sequencing coverage
- Read length
- Accuracy
- Multiple
- Control
- Replicates
- The number of target regions

Select the Sequencing Platform

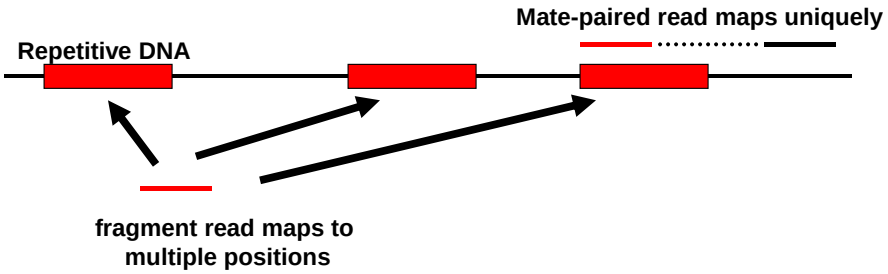


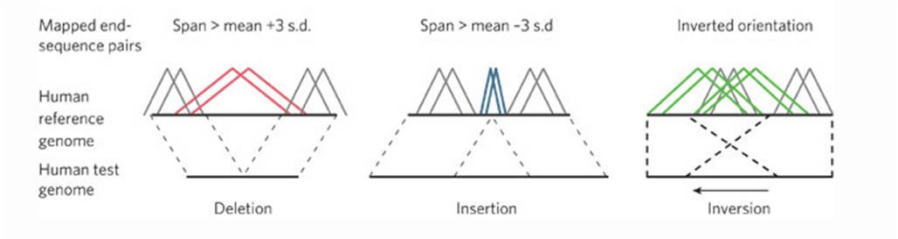
MacLean et al. Nat. Rev. Microbio 2009

Select Library Types for Different Applications



Advantage of Mate-pair/Paired-end reads





How Many Reads Do I Need to Survey the Transcriptome?

The number of reads needed is dictated by the complexity of application

Application	Complexity	Reads	Estimate mappable reads needed	Samples SOLID 3
Small RNA Discovery	Low	35bp	~10M	Up to 20/slide
SAGE	Low	35bp	5M	40/ slide
Expression of annotated genes	Mid	50bp	Minimum 50M (human	Up to 4/slide
Whole Transcriptome Discovery	High (alternative transcripts & splicing	50 bp	Minimum 100 million (human)	2/slide
Allele Specific Expression	High (variants to be defined)	50 bp	> 150 million (human)	1/slide

* Current best estimates from literature and internal research

applied biosystems part of Life technologies

Coverage Needed for SNP

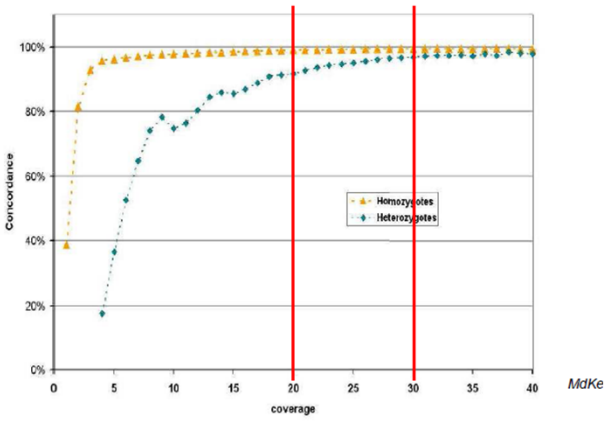


Figure 3. Dependence of genotype calling by dibase sequencing on depth of sequence coverage. The %dbSNP concordance is shown for homozygous and heterozygous SNPs at each level of coverage up to 40x. Two reads of the variant allele are required to call a homozygous SNP while two reads of each allele are required to call a heterozygous SNP.

McKernan K J et al. Genome Res. 2009

How to Compute the Sequencing Coverage

- **Genome coverage**
  - Raw genome coverage: mappable read * read length/genome size
- **Poisson distribution**

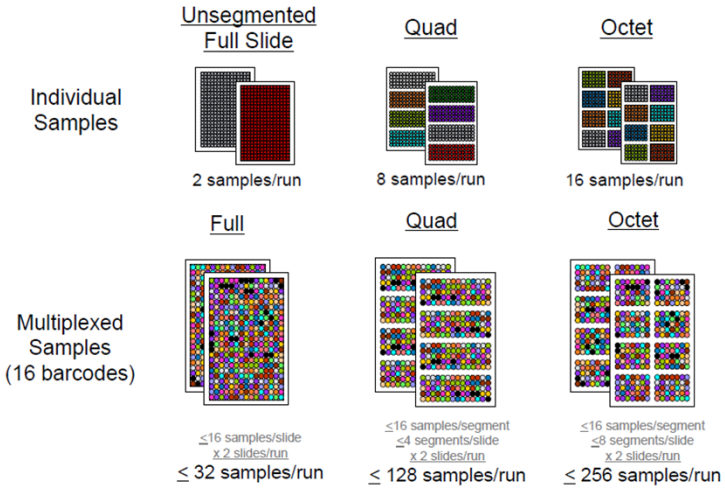
$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!},$$

lambda: average coverage

Probability of getting k reads from a base given the average coverage lambda

Select Format and Bar-code

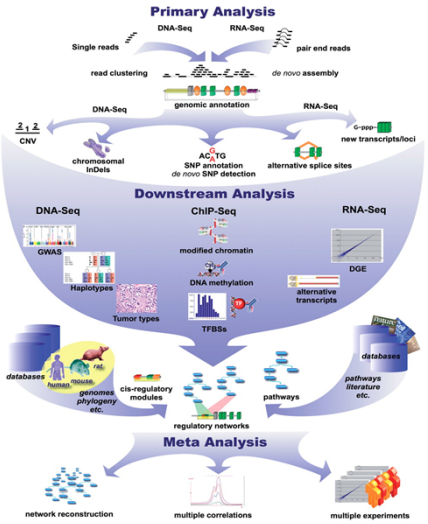
3 Slide Formats for Sequencing



Basic Data Analysis

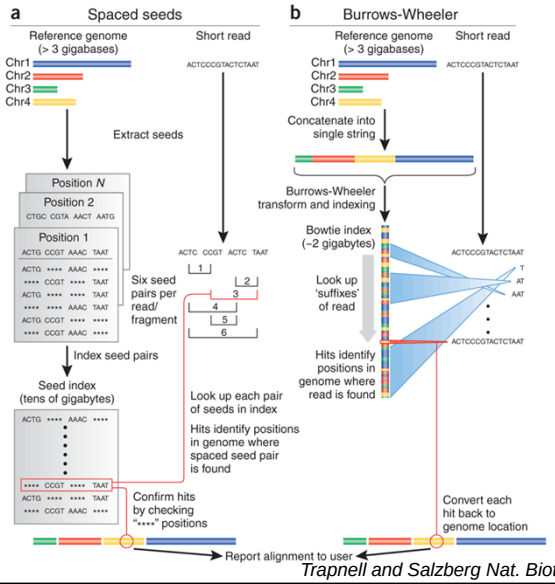
Overview of NGS-based Analysis Strategies

- **Initial analysis**
  - Mapping
  - Assembly
- **Application-specific analysis**



Werner Brief Bioinform 2010 11:499

How to Map Billions of Shot Reads onto Genomes



Trapnell and Salzberg Nat. Biotechnol. 2009 27: 455

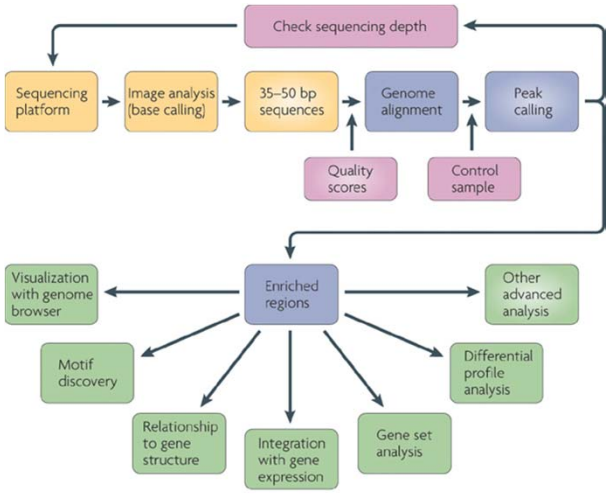
Popular Short-read Alignment Software

Program	Website	Algorithm	SOLiD	Long-read	Gapped	Pired-end	Open Source
Bfast	http://sourceforge.net/projects/bfast/	hashing ref.	Yes	No	Yes	Yes	Yes
Bowtie	http://bowtie.cbcb.umd.edu	FM-index	Yes	No	No	Yes	Yes
BWA	http://maq.sourceforge.net/bwa-man.shtml	FM-index	Yes	Yes	Yes	Yes	Yes
MAQ	http://maq.sourceforge.net	hashing reads	Yes	No	Yes	Yes	Yes
Mosaik	http://bioinformatics.bc.edu/marthlab/Mosaik	hashing ref.	Yes	Yes	Yes	Yes	No
Novoalign	http://www.novocraft.com	hashing ref.	No	No	Yes	Yes	No

Assembly Algorithms

- **ABySS**
  - http://www.bcgsc.ca/platform/bioinfo/software/abyss
- **Velvet**
  - Needs about 20-25x coverage
  - http://www.ebi.ac.uk/~zerbino/velvet/
- **AllPaths**
  - Requires 40x coverage
  - http://www.broadinstitute.org/science/programs/genome-biology/computational-rd/computational-research-and-development

Overview of ChIP–seq Analysis

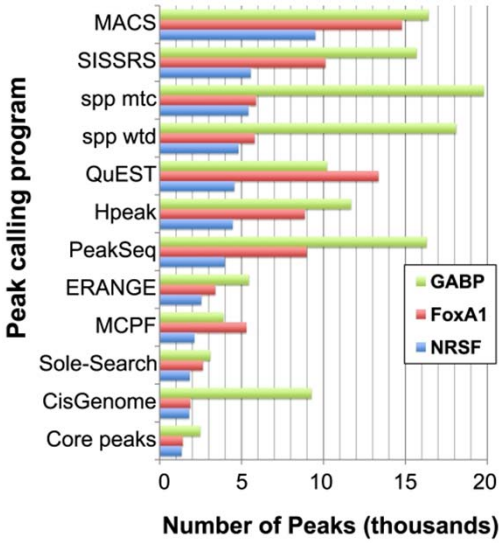


Nature Reviews | Genetics

Park Nat. Rev. Genet. 2009 10:669

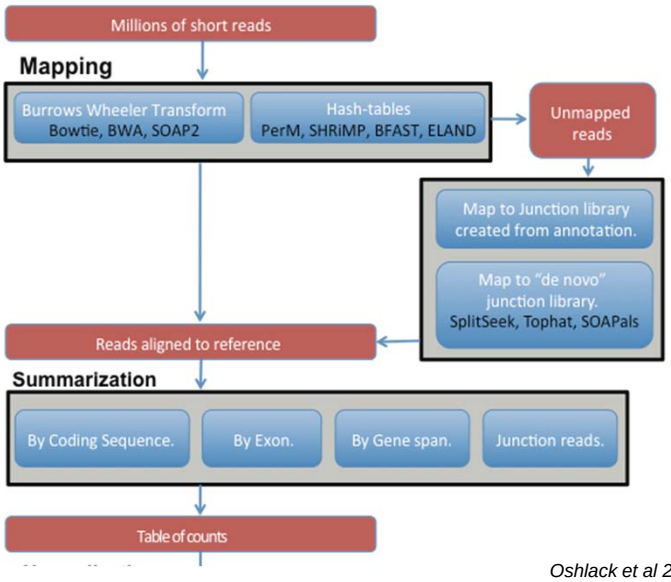
Peak Calling

- **Three categories**
  - TF binding site: a few hundred base pairs or less
  - RNA polymerase binding regions: up to a few kilobases
  - histone marks: up to several hundred kilobases
- **Histone modification**
  - ChIPDiff
  - ChromaSig



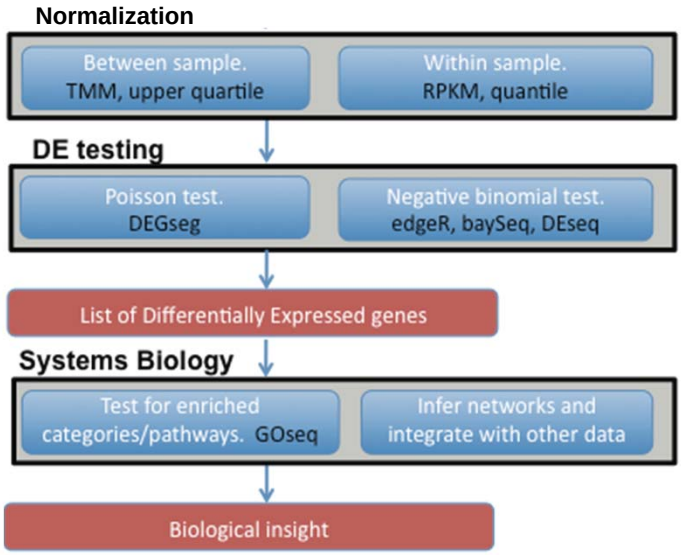
Wilbanks and Facciotti PLoS One 2010

Overview RNA-seq Analysis Pipeline for Detecting Differential Expression (Part I)



Oshlack et al 2010 Genome Biol.

Overview RNA-seq Analysis Pipeline for Detecting Differential Expression (Part II)

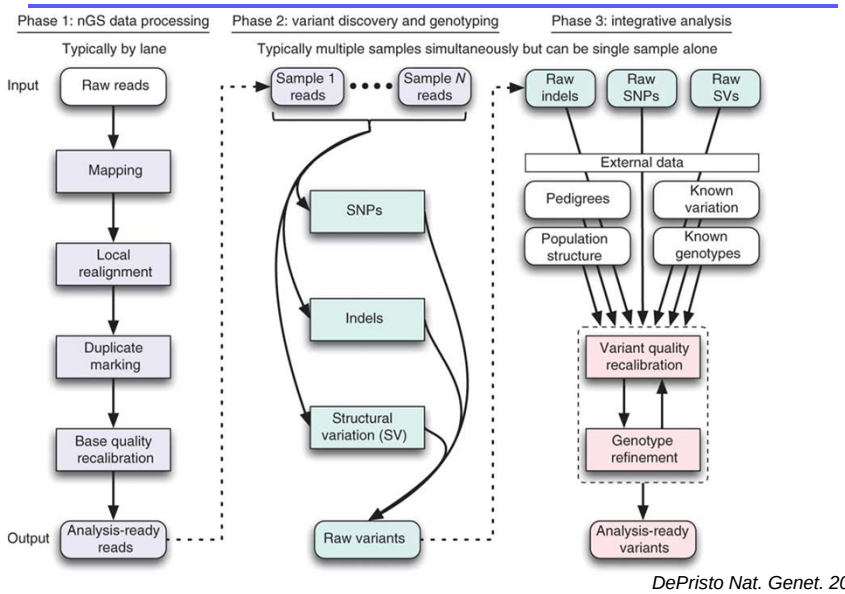


Oshlack et al 2010 Genome Biol.

Software and Tools for Differential Expression Analysis of RNA-seq

- **Junction Mapper**
  - SpliceMap: http://www.stanford.edu/group/wonglab/SpliceMap/
  - TopHat: http://tophat.cbcb.umd.edu/
  - G-Mo.R-Se: http://www.genoscope.cns.fr/externe/gmorse/
- **Summarization**
  - Cufflinks: http://cufflinks.cbcb.umd.edu/
  - ALEXA-seq: http://www.alexaplatform.org/alexa_seq/
- **Differential Expression**
  - BioConductor packages: edgeR, DEGseq, DESeq, and baySeq
- **Functional Analysis**
  - GOseq

Framework for Variation Discovery and Genotyping



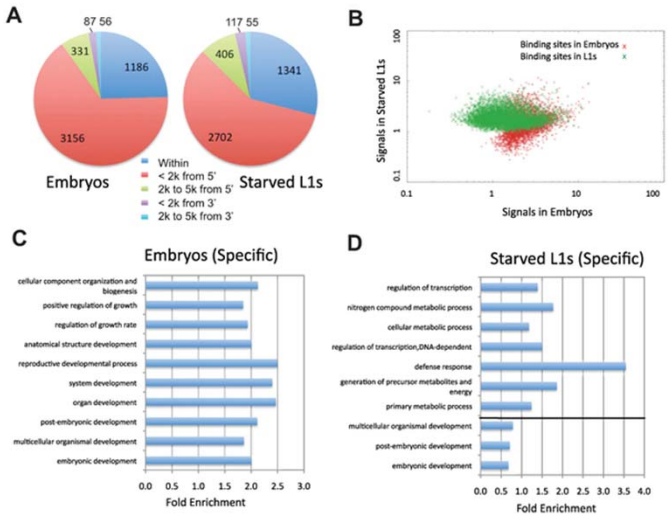
DePristo Nat. Genet. 2011

Computational Tools for Mutation Detection

- **Mutation calling**
  - DiBayes: LifeTech software
  - GATK: http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit
  - Samtools: http://samtools.sourceforge.net/
  - VarScan: http://varscan.sourceforge.net/
- **Indel calling**
  - Pindel: http://www.ebi.ac.uk/~kye/pindel/
- **Copy number analysis**
  - CBS: BioConductor package
  - SegSeq: http://www.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=182

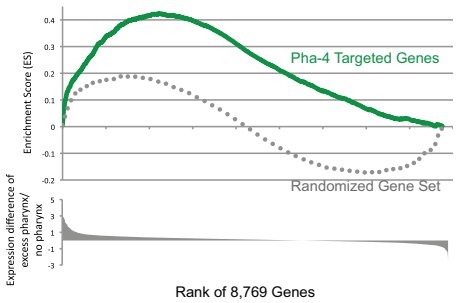
Advanced Data Analysis

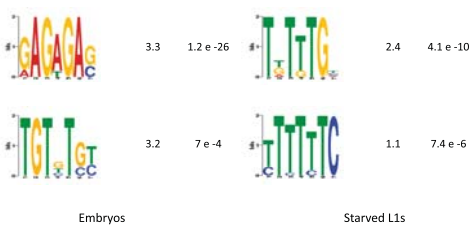
Characterization of TF Binding Patterns and Gene Targets



Zhong M et al. PLoS Genetics 2010

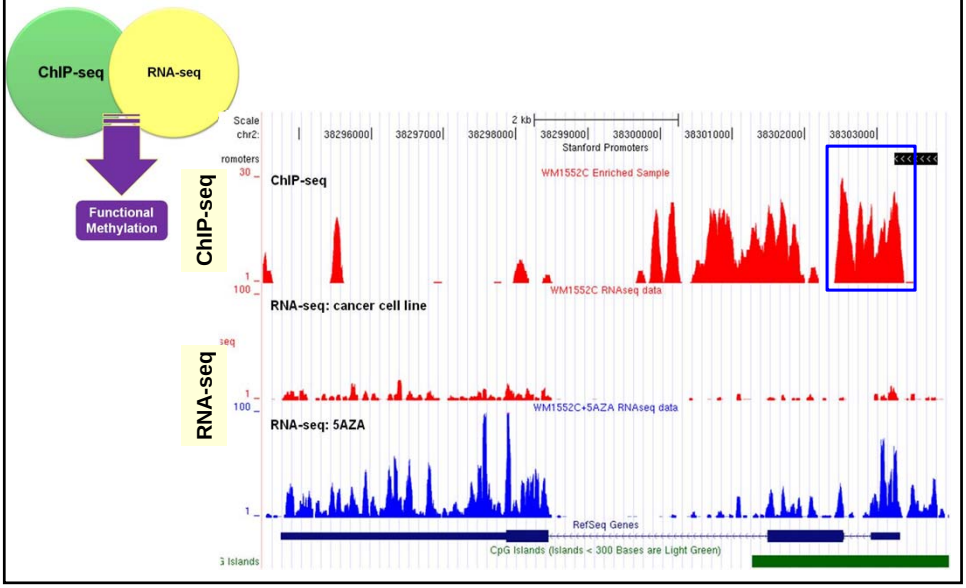
Characterization of TF Binding Patterns and Gene Targets (Cont.)



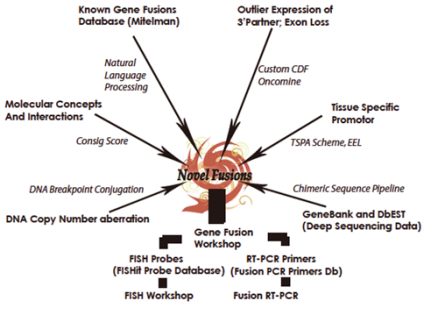


Zhong M et al. PLoS Genetics 2010

Integration of ChIP-seq and RNA-seq



An Integrative Approach to Reveal Driver Gene Fusion from Paired-end Sequencing Data in Cancer



- **Built a concept signature based on the associations of certain cancer characteristics between genes; altered biochemical pathways, molecular interactions, and functional annotation**
- **Applied concept signature to NGS**
- **Identify a gene fusion, R2HDM2-NFE2 in a lung cancer cell line (H1792)**

Useful Links

- **LifeTech**
  - http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing.html
- **SEQanswers**
  - http://seqanswers.com/
  - http://seqanswers.com/wiki/Software/list
- **Blog**
  - http://rna-seqblog.com/
  - http://mirnablog.com/

Shared Resources

- **Analytical Genomics Core (Lake Nona)**
  - http://intranet/researchsupport/sr/genomicsln/Pages/Home.aspx
- **Bioinformatics Shared Resource**
  - http://intranet/researchsupport/sr/bioinformaticsLJ/Pages/Home.aspx
  - La Jolla Campus
  - Building 10, Room 2045, 2046
- **Applied Bioinformatics Core (Lake Nona)**
  - http://intranet/researchsupport/sr/bioinformaticsLN/Pages/Home.aspx
  - Lake Nona Campus,
  - Room A2855